Capstone Project 2

Home Credit Default Risk

Shabnam Attaran

Outline

- Overview
- Business Objective
- Data
- Data Cleaning & Exploratory data analysis(EDA)
- Machine Learning Model
- Conclusion and future work

Overview

There are lots of people who do not particularly have a prior credit history, for example students, small businessmen, etc. who need credits, be it for studies, or for setting up some sort of businesses. Without adequate credit history, the lending organizations find it difficult to lend credits to such people, as these loans could be associated with high risks. In these kinds of situations, some lending organizations even tend to exploit the borrowers by asking for too high of an interest rate.

There are another subset of people, who do have prior credit history, which could be with the same organization or some other organizations. However, going through that historical data could be very time consuming and redundant. This would scale up even further as the number of applicants increases.

For such cases, if there could be a way through which the lending organization could predict or estimate the borrower's repayment capability, the process could be streamlined and be made effective for both the lender and the borrower. It could save resources both in terms of humans and time.

Business Objective

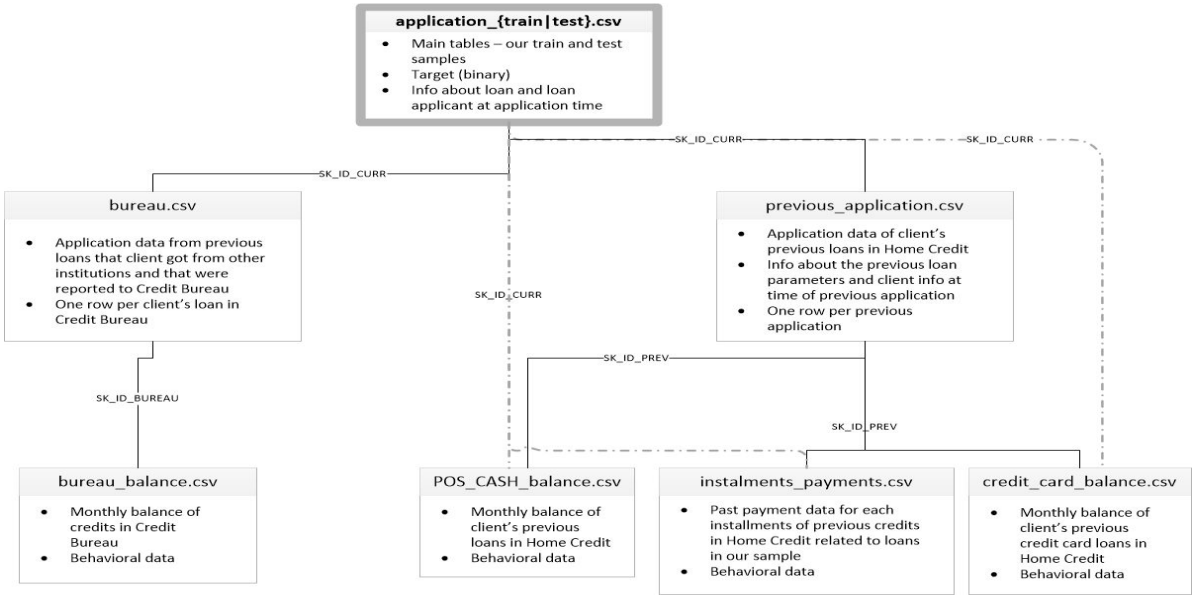
So the main two questions that the lender needs answer to are:

1) How risky is the borrower?

2) Given the borrower’s risk, should we give him/her loan?

Data

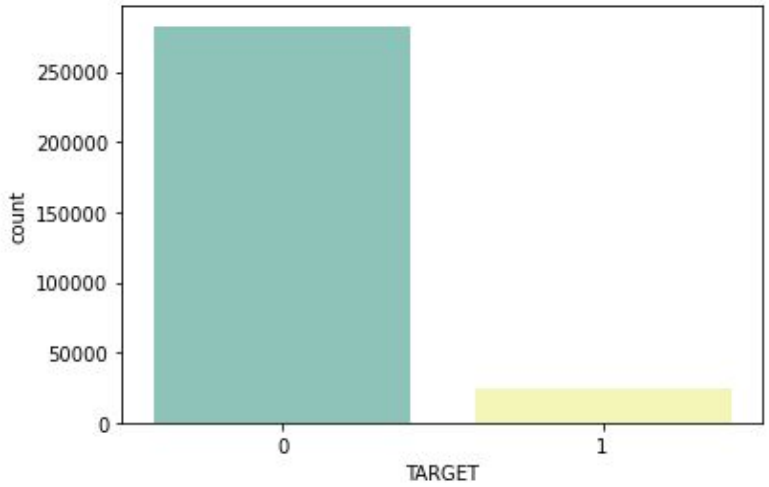
I got the data from kaggle competition



Data Cleaning

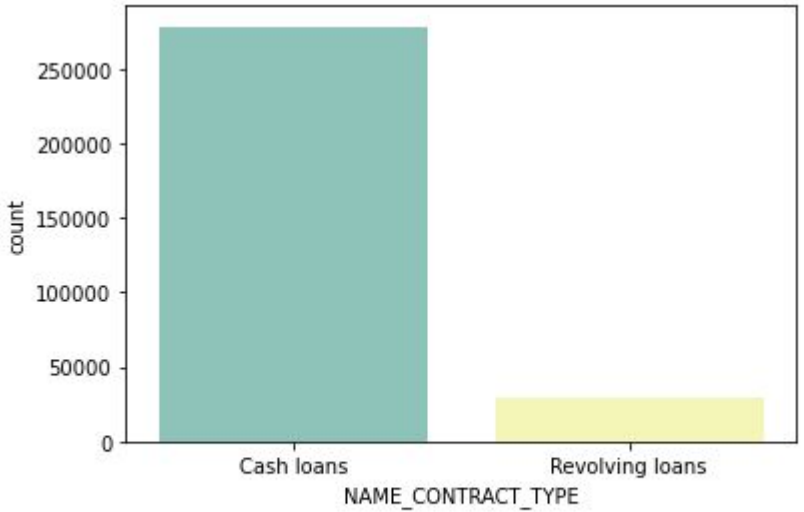
- Categorical features(one hot encoding)
- Aligning training and test data
- Missing values

EDA



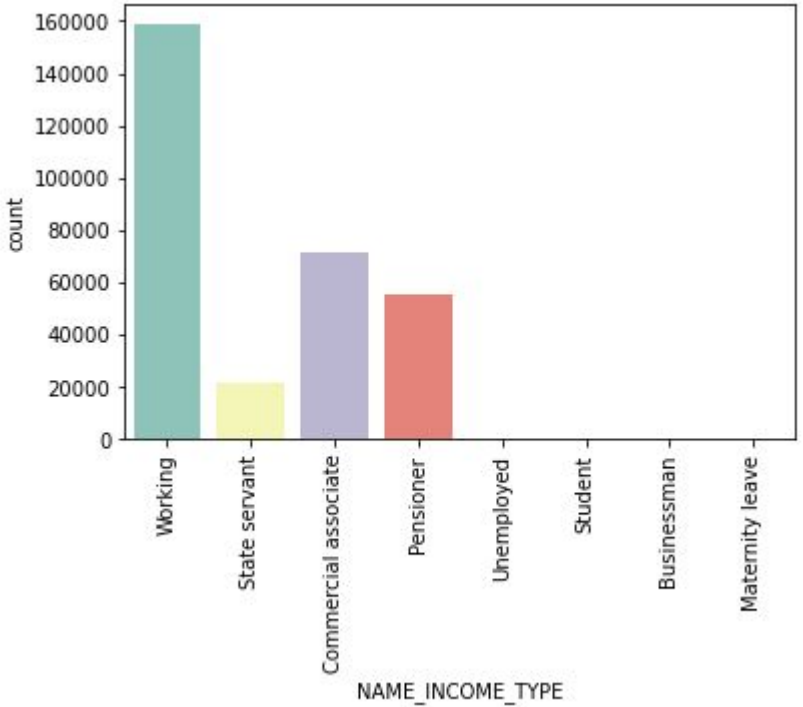
Target is very unbalaced which is normal. Which means most peopel are cleared to get a loan

EDA

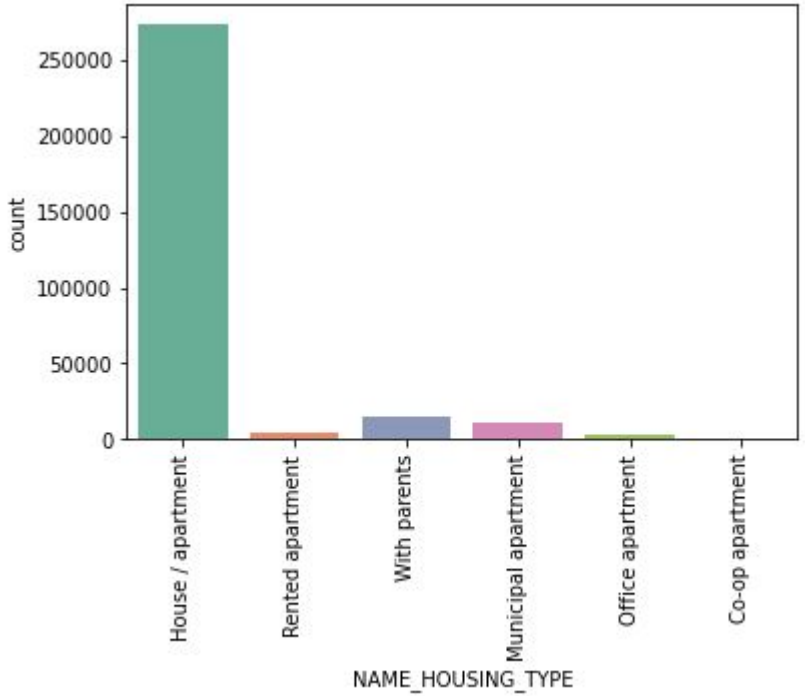


Around 90% of contract type is cash loan and 10% is revolving loan

EDA


160000
140000
120000
100000
80000
60000
40000
20000
0
count
Working
State servant
Commercial associate
Pensioner
Unemployed
Student
Businessman
Maternity leave
NAME_INCOME_TYPE

EDA


250000
200000
150000
100000
50000
0
count
House / apartment
Rented apartment
With parents
Municipal apartment
Office apartment
Co-op apartment
NAME_HOUSING_TYPE

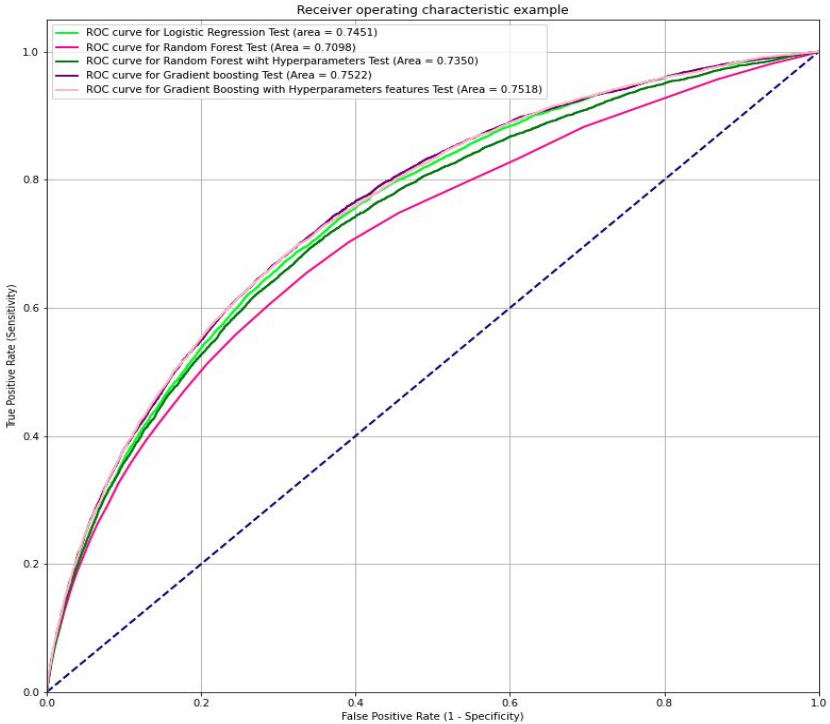
Classification

- Supervised: The labels are included in the training data and the goal is to train a model to learn to predict the labels from the features
- Classification: The label is a binary variable, 0 (will repay loan on time), 1 (will have difficulty repaying loan)

Machine Learning Model

- Logistic Regression
- Random Forest
- Random Forest with hyperparameter
- Gradient Boosting
- Gradient Boosting with hyperparameter

Comparing different Models



Conclusion and Future work

In this project, I used the kaggle machine learning competition. First I made sure I understand the data and the problem I am trying to solve. For this I performed some EDA to find some trends and relationships in the data. After that I did some preprocessing like encoding categorical features, imputation and scaling.

In this project I only used the main csv file . For future work we can use all of the csv files we have to get a better model.

We can also improve the models by doing more feature engineering in the future.